HW 4

Zoe Werner

10/28/2024

This homework is designed to give you practice working with statistical/philosophical measures of fairness.

The paper linked below[1] discusses potential algorithmic bias in the context of credit. In particular, banks are now regularly using machine learning algorithms to do an initial screening for credit-worthy loan applicants. In section 4.5.2, this paper reports the rates at which various racial groups were granted a mortgage. If we assume that it is a classifier making these predictions[2] what additional information would be necessary to assess this classifier according to equalized odds?

Student Input. We would need to determine the confusion matrix for each racial group we are comparing and use the true positive data and false positive data from each group to calculate a true positive rate and false positive rate. If the number resulting from either the true positive rate or the false positive rate is higher then our choice of epsilon (or the legal precedent of 0.2), it displays an unfair distribution of error, proving bias in the model. If either of these standards are violated, equalized odds is violated and the classifier does not pass the fairness test.

Show or argue that the impossibility result discussed in class does not hold when our two fringe cases[3] are met.

Student Input The incompleteness theorem states that all three fairness criteria cannot be met except in the case of two fringe cases being met. When the two fringe cases are met, the impossibility result of the incompleteness theorem does not hold because the separation (states that the model has equal rates among groups) and sufficiency (states that the model predicts accurately for all groups with a bit of variability allowed) prove independence (any discrepancies in the data are unrelated to the group). If there are equal rates and very similar accuracy for each group, the errors are not attributed to racial or social factors, proving that all three fairness criteria are met.

How would Rawls's Veil of Ignorance define a protected class? Further, imagine that we preprocessed data by removing this protected variable from consideration before training out algorithm. How could this variable make its way into our interpretation of results nonetheless?

[1] https://link.springer.com/article/10.1007/s00146-023-01676-3

[2] It is unclear whether this is an algorithm producing these predictions or human

[3] a) perfect predicting classifier and b) perfectly equal proportions of ground truth class labels across the protected variable

Student Input Rawls Veil of Ignorance would define a protected class as the social standing, race, gender, or other sensitive factors of the disparate group. He argued that if we were to step behind a veil and imagine ourselves in the persons position without knowing who they are, our decision will be impartial and not affected by the persons sensitive factors - such as social standing, gender, or race. The Veil of Ignorance refers to "raising your floor rather than raising your ceiling," and "walking a mile in someone elses shoes." In essence, it allows the decider to look past certain aspects of moral/social concern and focus on decision making for the greater good. By removing this variable before training our algorithm, it may lead to more impartial algorithmic results because it shields the algorithm from potentially biasing information, but the interpretation could introduce it back into consideration. Although the algorithm is trained with Veil of Ignorance standards, the interpretation will be impacted by the interpreters personal bias. The definition of a vulnerable group is not explicitly revealed by Rawl, so the interpreter may feel that the algorithm is overly sensitive or not sensitive enough, leading to its reintroduction into the result interpretations. Along these lines, if we removed the protected class originally, the decision of what we define as vulnerable is up to our discretion, potentially excluding certain factors that others feel should be considered, or vice versa. If, as a result, there are disparities in the data, it will work in favor of the most protected class and overtly against the least protected class. Therefore, any class that is not protected will still be impacted.

Based on all arguments discussed in class, is the use of COMPAS to supplement a judge's discretion justifiable. Defend your position. This defense should appeal to statistical and philosophical measures of fairness as well as one of our original moral frameworks from the beginning of the course. Your response should be no more than a paragraph in length.

Student Input The use of COMPAS to supplement a judge's discretion is not justifiable. Based on the moral framework of Consequentialism, an action is justified by the outcome/consequences and on the subset Utilitarianism, which states that correct actions minimize pain and maximize pleasure. The bias-impacted algorithm has negative consequences and maximized pain, rather then pleasure, for the convicts being considered for parole. Potentially, it could have negative consequences on the judges that relied on COMPAS, based on the judges own moral convictions relating to racism. In addition, the costs of using the algorithm as a supplemental decider outweigh the consequences. The positive outcomes include a mathematical evaluation that judges can use to support their decisions, the ability to predict potential re-offenders and prevent parole as a result. The costs include unfair denial of parole based on race, potential mathematical confirmation of a judges racist tendencies, and inaccurate results (approximately only 3/4 accuracy). Based on the measures of fairness, the algorithm is unjust. Based on our calculations in class, COMPAS violated the separation criteria based on the equalized odds test, which is violated in either the false positive rate or the true positive rate are greater then epsilon. While the false positive rate was not violated, the rate of true positives between black and white were convicts were above the legal precedent of epsilon, which is 0.2. In addition, COMPAS violated the independence criteria because it was proven to have disparate impact and violated statistical parity, which proved black convicts had a larger, unequal (in relation to white convicts) chance of being classified as re-offenders.